Topic Mining on IRCTC App Reviews

Aryan Agarwal, 241110012, aryana24@iitk.ac.in
Harsh Baid, 241110026, bharsh24@iitk.ac.in
Vraj Patel, 241110080, vrajb24@iitk.ac.in
Aditya Azad, 241110002, adityaazad24@iitk.ac.in

CS685: Data Mining

Abstract

This project discusses various methods identifying key topics in the reviews left by users on IRCTC app. We have gathered reviews publicly available from Google Play Store for analysis purposes. Project aims to highlight strengths and weaknesses of App expressed in the reviews. This analysis provides App developers with actionable insights and enable them to make data-driven decisions. This problem comes under unsupervised learning. That means identifying and clustering patterns in text without prior labels. Project includes various natural language processing techniques to mine topics from review data. The process comprises of several stage including data collection, pre-processing and multiple iterations of model refining. As a result we were able to find around 66 unique topics in the data using BERT embeddings and HDBSCAN clustering on the data transformed using a UMAP model. We built a hierarchy of these topics to be able to visualize easily the data and topics extracted, and make corresponding analysis and suggestions for how IRCTC can use these insights for improving their app workflow.

Contents

1 Motivation of the Problem

1.1 Background

The IRCTC app serves millions of users for booking train tickets and accessing other railway services in India. As such, understanding user sentiment and feedback is crucial for continuous improvement. Analyzing user reviews on platforms like the Google Play Store can provide valuable insights into user satisfaction and highlight areas needing attention.

1.2 Problem Statement

Extracting key topics from IRCTC app reviews to identify strengths and weaknesses as perceived by users.

1.3 Aims and Objectives

Aim: The primary aim of this project is to equip IRCTC with a comprehensive understanding of user sentiment and feedback trends, enabling data-driven improvements to the app.

Objectives:

- To gather and process a large dataset of user reviews from the Google Play Store.
- To employ NLP techniques to clean and preprocess text data for effective topic extraction.
- To apply clustering techniques and embeddings for topic identification, highlighting key themes in user feedback.
- To summarize the findings in a way that aids IRCTC in decision-making for app enhancements.

1.4 Methodology Overview

The project follows a multi-step methodology to mine insights from user reviews:

- **Data Collection:** We began by scraping review data from the Google Play Store using a custom script. This script extracted essential fields such as review content, rating, and timestamps, creating a dataset that represents user feedback over time.
- **Text Cleaning and Preprocessing:** Extensive text preprocessing was applied, including lemmatization, stop-word removal, and translation of mixed-language content into English. A custom cleaning function was developed and refined to retain meaningful text tokens.
- **Topic Mining:** We tried multiple approaches for this- including using LDA (Latent Dirchelet Allocation), Clustering based mining and using pre-existing embedding models. Our final approach chosen after all analysis/trials is to use BERT embeddings combined with HDBSCAN clustering and UMAP dimensionality reduction.
- **Hierarchy of Topics:** We organized the extracted topics into a hierarchy to visually represent relationships between themes, enabling the IRCTC team to quickly identify major areas of concern and user appreciation.

1.5 Organization of the Report

This report is organized into seven chapters:

- Chapter 1 provides an introduction, including the project overview, aims, objectives, and methodology.
- Chapter 2 describes the data collection and preprocessing steps.
- Chapter 3 presents the various topic mining approaches applied in this project.
- Chapter 4 covers the sentiment analysis performed on the collected reviews.
- Chapter 5 discusses the key findings, insights, and recommendations derived from the topic analysis.

2 Data Used

2.1 Scraping Text Reviews

We scraped approximately 100,000 reviews of the IRCTC android app from Google Play Store providing us the raw data needed to identify key topics.

Following features were collected for each data point:

- **Review Content:** The main body of user feedback.
- **Rating:** A numerical rating provided by users in form of stars. Rating can be anywhere from 0 stars to 5 stars where higher number of stars signifies better rating.
- **Timestamp:** The date when the review was posted.
- **Other:** Other components of review which includes review ID and review No.

2.2 Scraping Methodology

We created a Python script that interacts with the public Google Play Store web page using selenium and scrapes user reviews to obtain raw review data.

2.2.1 Data Extraction

The scraper script was made using python libraries named selenium and beautiful soup. The script was designed to cycle through multiple pages of reviews on Google PlayStore website and obtain latest reviews to extract up to date data.

2.2.2 Error Handling

Minimal Error Handling was required at this stage. Most of it was done manually.

2.2.3 Data Storage

Scrapped reviews were stored in CSV format for easy handling and processing.

2.3 Sample Data and Screenshots

Screenshots of working scrapper (Figure 1) and collected sample data (Figure 2) are given below.

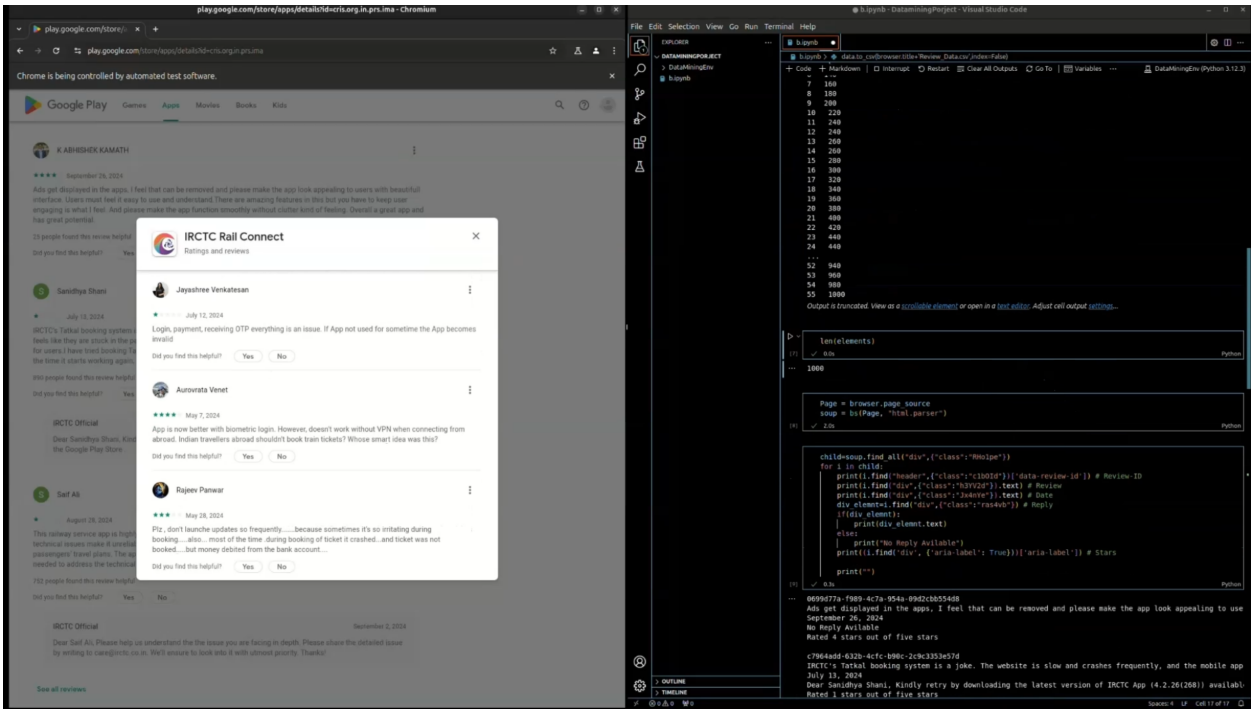

Figure 1: Screenshot of the scraping process.

```
data.head()
```

[16] ✓ 0.0s Python

	Review_No	Review_ID	Review_Date	Review_Star	Review	Reply
0	1	0699d77a-f989-4c7a-954a-09d2cbb554d8	September 26, 2024	Rated 4 stars out of five stars	Ads get displayed in the apps, I feel that can...	NA
1	2	c7964add-632b-4cfc-b90c-2c9c3353e57d	July 13, 2024	Rated 1 stars out of five stars	IRCTC's Tatkal booking system is a joke. The w...	Dear Sanidhya Shani, Kindly retry by downloadi...
2	3	5eb728c4-88b6-4d9d-ae3b-4187293f3b06	August 28, 2024	Rated 1 stars out of five stars	This railway service app is highly disappointi...	Dear Saif Ali,\nPlease help us understand the ...
3	4	c61bda54-a63c-4ab3-a0fc-6049e03284cd	September 24, 2024	Rated 1 stars out of five stars	Third class app!😡Not at all worth! 1)I was boo...	NA
4	5	86c806f8-089f-48c1-94d1-aff3a884c0ba	September 26, 2024	Rated 1 stars out of five stars	1. App doesn't need ads. One ad covered captch...	NA

Figure 2: Screenshot of a sample of the collected data.

2.4 Demo Video

In addition to images, we have also attached a demo video of working scrapper for reference. Please click on the following link to view the demo video: Scraping Demo Video We are also submitting this video in our zip file submission for immediate reference.

3 Methodology

3.1 Overview

We tried multiple approaches to extract reasonable topics from the data listed sequentially in the report. We learned certain inferences and ideas about the topics from each approach, which helped us model the following approach better.

3.2 Approach Evaluation Metrics

Since this project is based on unsupervised learning, we lacked labeled data for straightforward model validation. We used a mix of metrics to compare the effectiveness of each approach:

- We used **Perplexity** as the deterministic metric to compare our approaches and model.
- We used metrics such as **Visual Plots** of intertopic distances between the topic clusters or t-SNE plots and human-based checking of critical topic reviews by checking key words in that topic distribution/cluster and compare it with actual sample reviews' context.
- We used **Domain Knowledge** to analyze the model performance, which was crucial in this project and, in general, any unsupervised learning problem.

3.3 Common pre-processing steps across Approaches

In each approach, we have explained how we performed text cleaning and transformation, if applicable. We performed some common preprocessing in each of the approaches listed below:

- Translating a mix of non-English languages to English using Google Translate
- The custom text cleaning function was refined over multiple iterations and analyses. This process involved converting text to lowercase, removing punctuation, performing lemmatization, and removing stop-words. We also handled cleaning of domain jargon terms as well based on our learnings from diffferent iterations

3.4 Visualization of Preprocessing

To evaluate the preprocessing steps' effectiveness, we generated the following visualizations:

3.4.1 Word Frequency Plot:

Domain-specific stop-words identified through domain knowledge and frequency plots were removed, as shown in the figure below. We can also evaluate the sparsity (99.41%) of the DTM (Document term Matrix) from the count vectorizer used to make this plot. It was low, thus giving us confidence that data is sufficiently sparse as needed.

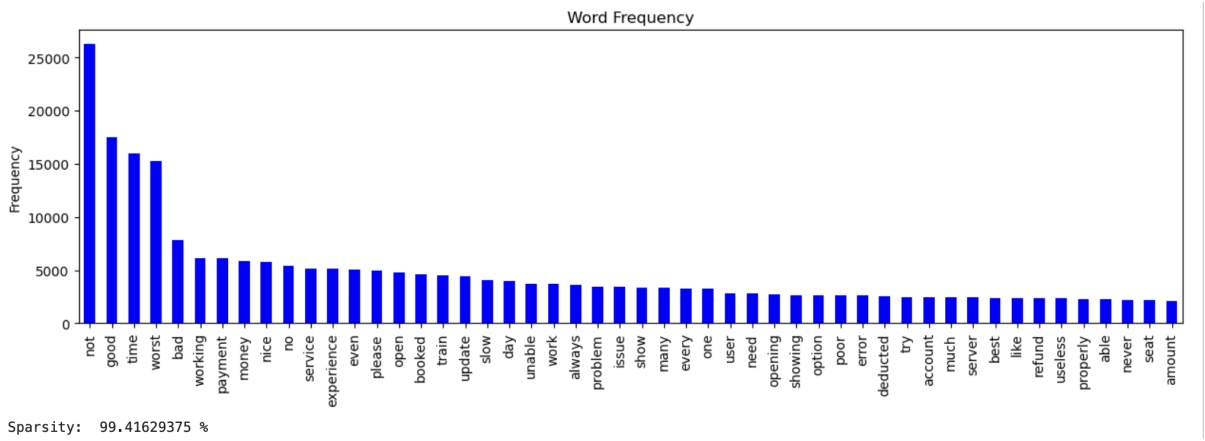

Figure 3: Word Frequency Plot showing the top 50 words after cleaning.

3.4.2 TF-IDF Frequency Plot

We also created a frequency plot using a TF-IDF vectorizer to examine the data distribution.

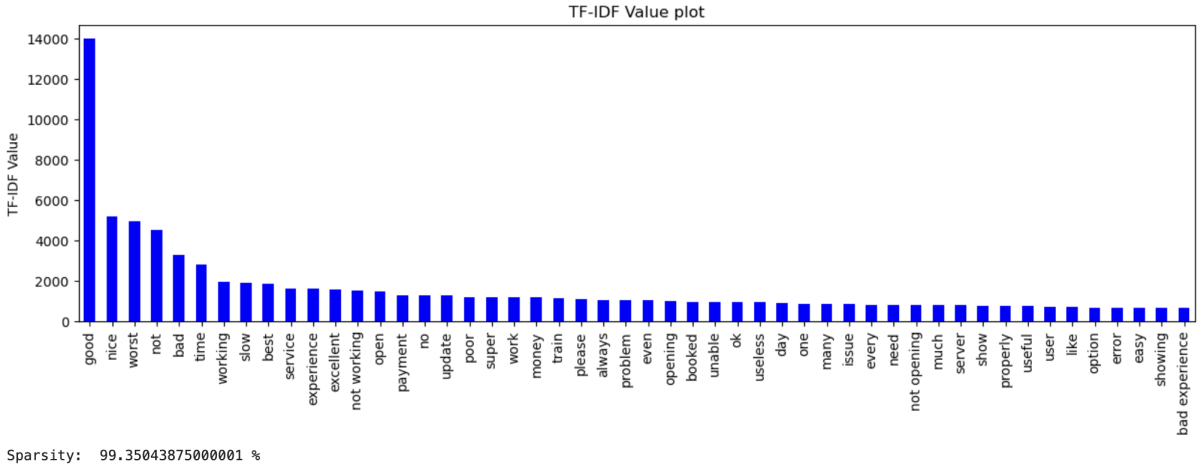

Figure 4: TF-IDF Value Plot showing term significance across documents.

3.4.3 Word Cloud

We also plotted a word cloud of the final words in the corpus to visualize any data better, identify keywords and outlier words, and act accordingly to handle them. We show the final plots after refining the corpus and handling domain jargon.

Figure 5: Word Cloud of the most frequent terms after preprocessing.

3.5 Approach 1: Latent Dirichlet Allocation (LDA)

We experimented with LDA with two vectorization methods, Count Vectorizer, and TF-IDF Vectorizer, and tuned each separately for optimal model performance.

3.5.1 Text Preprocessing

We performed basic text cleaning on the translated data, and vectorized the data with mentioned vectorization techniques respectively. With this, our Document Term Matrix (hereon referred to as DTM) was ready for modeling.

3.5.2 Modeling

For modeling, we performed hyperparameter tuning for n_components (i.e., number of topics), learning decay, and maximum iterations. For both TF-IDF and Count Vectorizer, we plotted different scores we got, as shown below. Modeling involved multiple iterations over fixing cleaning functions and trying combinations of vectorization and modeling hyperparameters each, to arrive at a suitable model.

Finalized Hyperparameters:

- Vectorizer: $ngram\_range = (1, 1)$, $min\_df = 4$, $max\_features = 800$
- LDA: $max\_iter = 10$, $learning\_offset = 20$, $n\_components = 10$

For a more detailed examination of the tuning process, refer to the Jupyter Notebook included with this report.

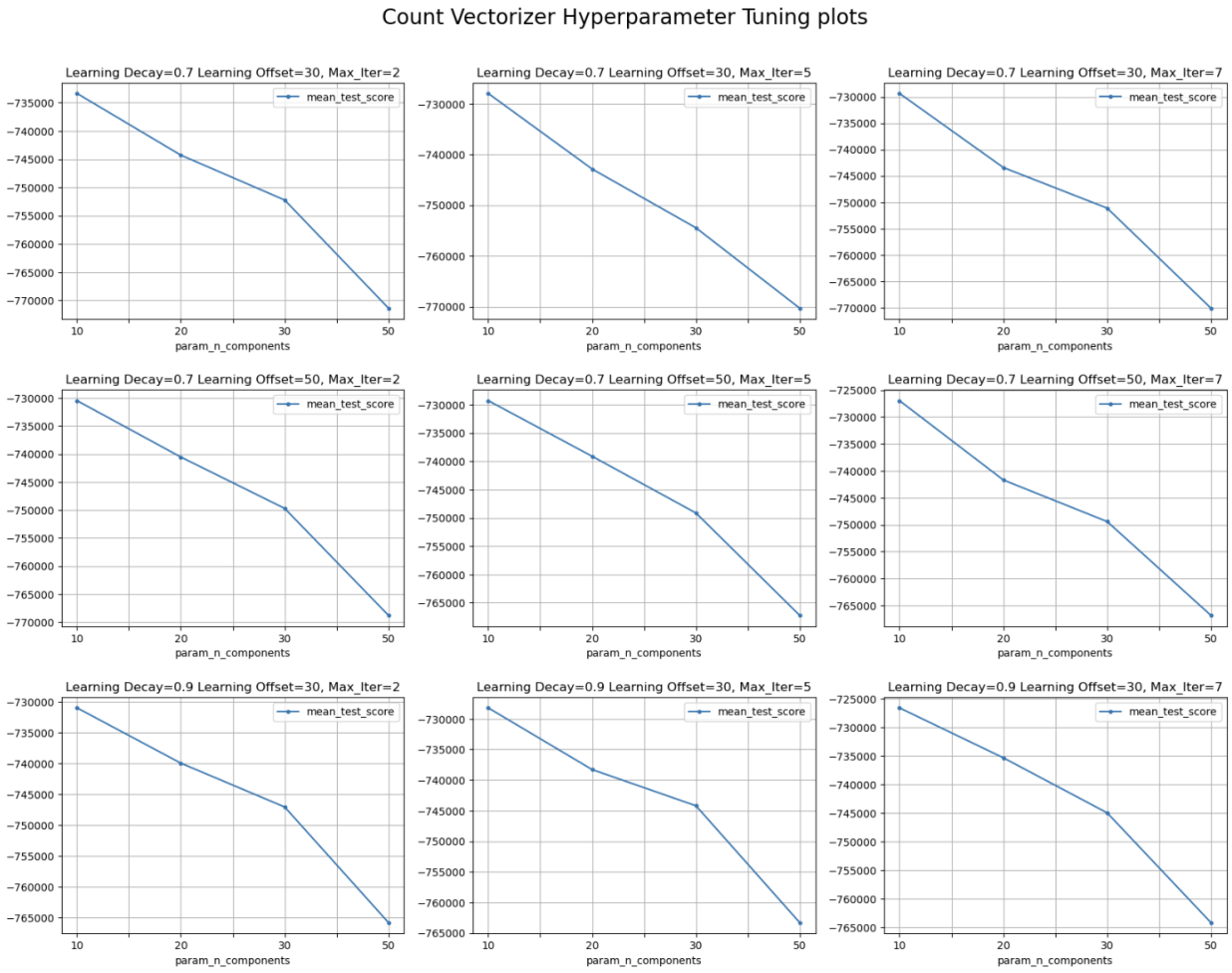

Figure 6: Hyperparameter tuning plot for Count Vectorizer.

3.5.3 Model Evaluation

- **Count Vectorizer Results:**
  - Log Likelihood: -2998005.574002954
  - Perplexity: 361.9337833598004
- **TF-IDF Vectorizer Results:**
  - Log Likelihood: -1168575.5252102495
  - Perplexity: 649.9993224239983
- **Intertopic Distance Plots Results:**
  - For Count Vectorizer Intertopic Distance Plot for Count Vectorizer.
  - For TF-IDF vectorizer Intertopic Distance Plot for TF-IDF Vectorizer.

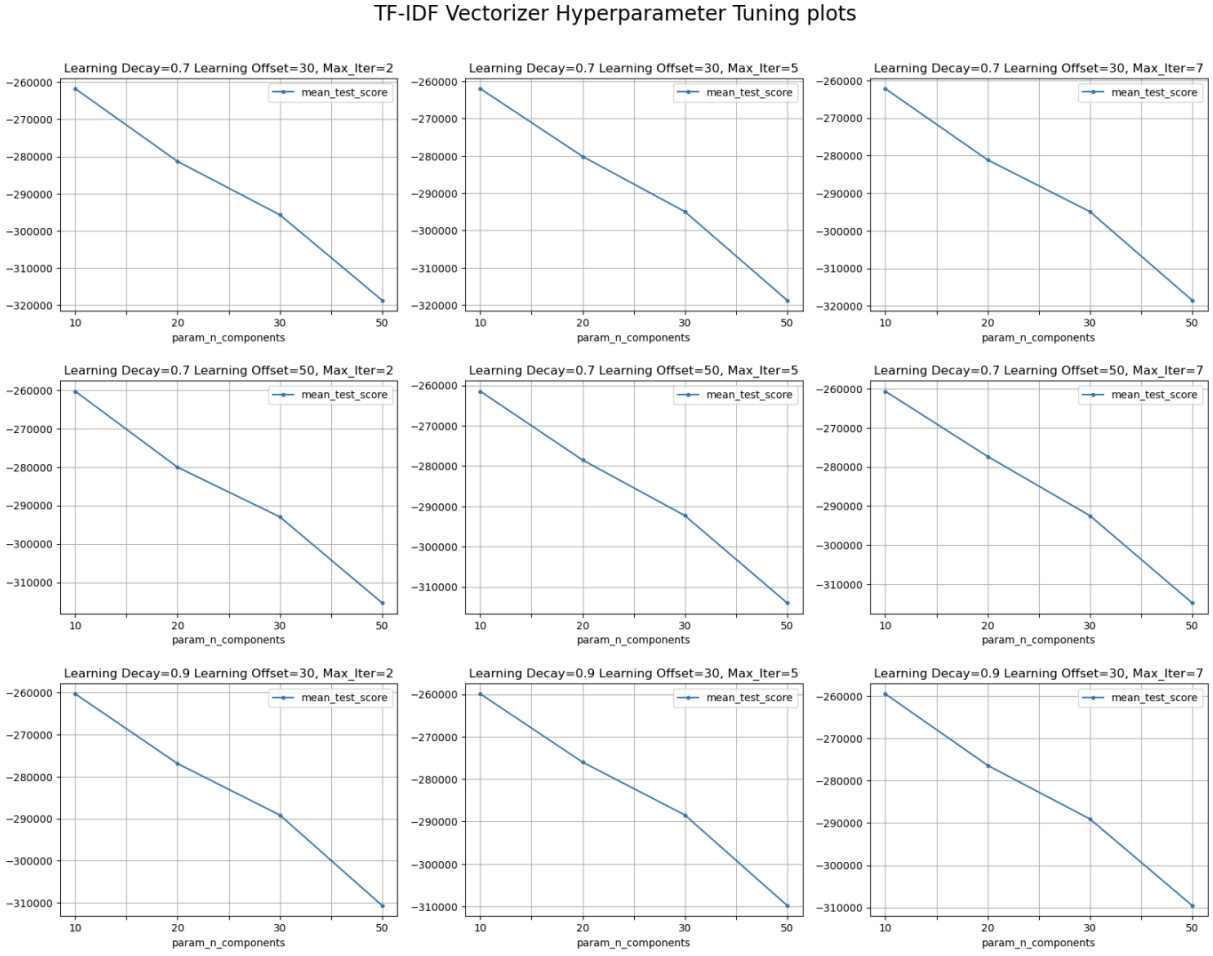

Figure 7: Hyperparameter tuning plot for TF-IDF Vectorizer.

3.5.4 Visual Outputs

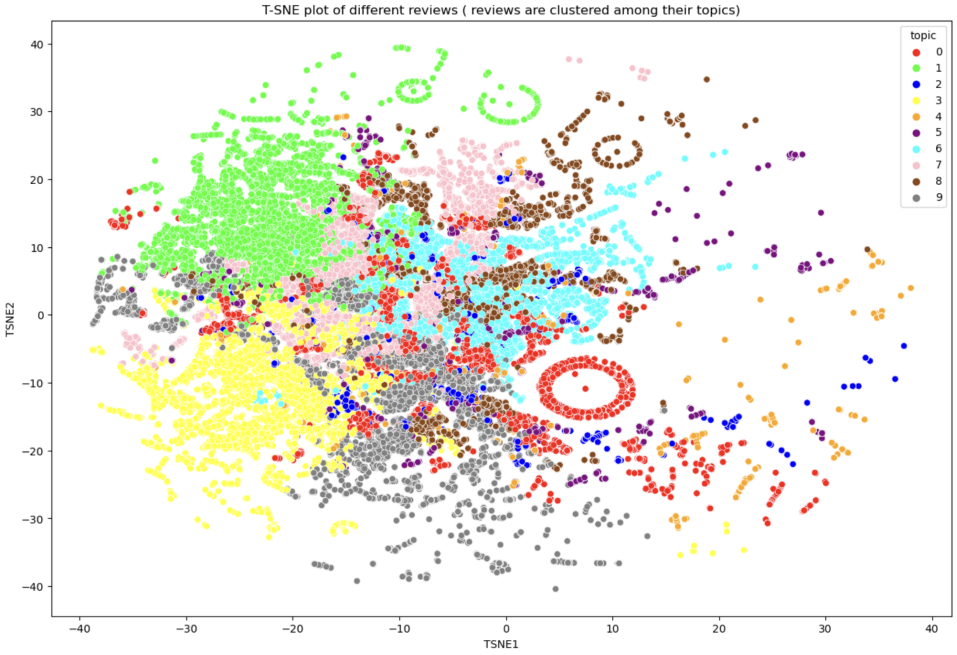

Figure 8: t-SNE plot for Count Vectorizer topics.

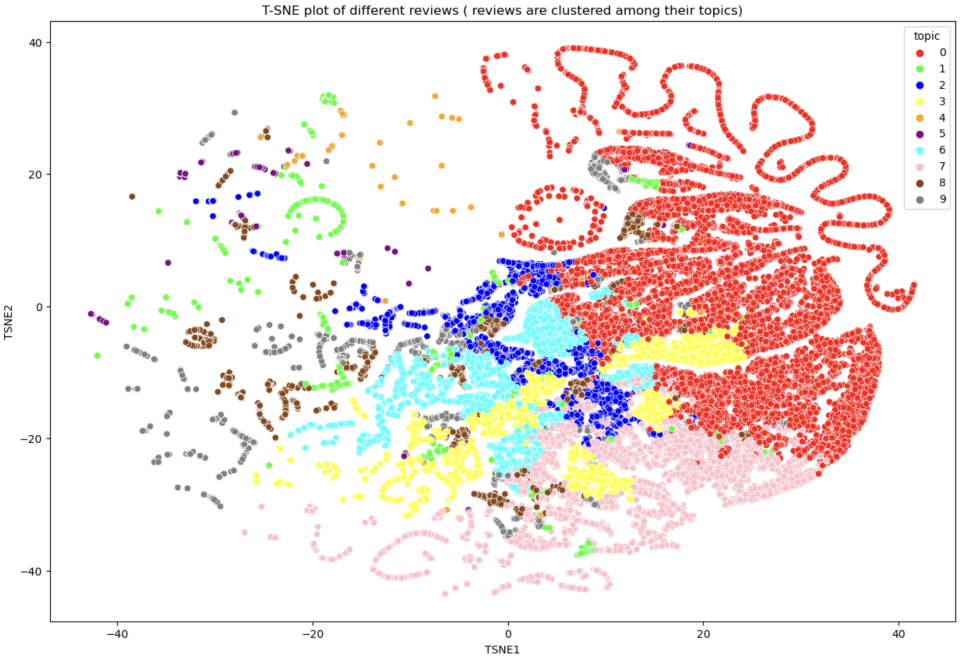

Figure 9: t-SNE plot for TF-IDF Vectorizer topics.

```
Topic 0:
[('poor', '1070.05%'), ('money', '852.35%'), ('ticket', '805.49%'), ('booked', '775.23%'), ('ticket booked', '476.54%'), ('deducted', '436.1
4%'), ('customer', '343.67%'), ('payment', '331.57%'), ('service', '322.4%'), ('money deducted', '318.43%')]
Topic 1:
[('train', '427.08%'), ('used', '336.03%'), ('superb', '283.49%'), ('train ticket', '241.56%'), ('worst used', '171.98%'), ('love', '159.1
3%'), ('bakwas', '150.45%'), ('bed', '149.56%'), ('responding', '134.0%'), ('book train', '97.35%')]
Topic 2:
[('login', '894.12%'), ('time', '635.4%'), ('every', '631.46%'), ('every time', '518.51%'), ('update', '441.54%'), ('pathetic', '411.79%'),
('password', '410.21%'), ('many time', '381.53%'), ('wrost', '358.95%'), ('account', '345.1%')]
Topic 3:
[('ad', '484.01%'), ('class', '348.87%'), ('hang', '311.64%'), ('add', '296.93%'), ('third', '256.83%'), ('star', '225.11%'), ('reservation',
'214.49%'), ('always', '213.6%'), ('third class', '202.61%'), ('late', '193.51%')]
Topic 4:
[('service', '812.54%'), ('railway', '578.97%'), ('good service', '381.03%'), ('doe', '305.8%'), ('full', '264.73%'), ('good', '253.77%'),
('help', '161.31%'), ('charge', '120.86%'), ('public', '100.12%'), ('doe work', '97.04%')]
```

Figure 10: Intertopic distance plot for Count Vectorizer.

```
Topic 0:
[('time', '1758.65%'), ('not', '1519.58%'), ('update', '1011.06%'), ('problem', '994.33%'), ('opening', '908.5%'), ('unable', '865.45%'), ('a
lways', '802.77%'), ('worst', '795.64%'), ('please', '786.41%'), ('every', '777.01%')]
Topic 1:
[('worst', '3735.26%'), ('service', '1557.43%'), ('super', '1260.29%'), ('seen', '424.48%'), ('one', '350.39%'), ('india', '281.08%'), ('use
d', '266.83%'), ('difficult', '226.2%'), ('happy', '211.49%'), ('life', '151.82%')]
Topic 2:
[('no', '542.47%'), ('train', '465.38%'), ('class', '427.51%'), ('option', '425.29%'), ('like', '378.73%'), ('add', '352.01%'), ('third', '29
6.04%'), ('superb', '291.45%'), ('reservation', '218.4%'), ('change', '208.68%')]
Topic 3:
[('slow', '1650.57%'), ('money', '1035.73%'), ('booked', '782.82%'), ('deducted', '722.96%'), ('not', '696.69%'), ('refund', '634.92%'), ('pa
yment', '551.37%'), ('amount', '538.67%'), ('better', '461.29%'), ('account', '444.11%')]
Topic 4:
[('good', '15173.43%'), ('fast', '284.49%'), ('operate', '61.04%'), ('issued', '42.11%'), ('service', '0.1%'), ('job', '0.1%'), ('easy', '0.
1%'), ('experience', '0.1%'), ('free', '0.1%'), ('overall', '0.1%')]
Topic 5:
[('nice', '5706.17%'), ('best', '1983.75%'), ('easy', '702.53%'), ('great', '521.08%'), ('connection', '105.93%'), ('train', '87.35%'), ('tra
sh', '57.99%'), ('hopeless', '51.23%'), ('forever', '42.04%'), ('avoid', '29.59%')]
Topic 6:
[('user', '642.32%'), ('not', '523.11%'), ('password', '442.99%'), ('friendly', '416.61%'), ('able', '357.29%'), ('id', '325.39%'), ('new',
'310.64%'), ('never', '307.83%'), ('download', '298.05%'), ('updated', '251.4%')]
Topic 7:
[('poor', '1189.99%'), ('useless', '715.28%'), ('waste', '575.91%'), ('helpful', '491.14%'), ('ad', '469.22%'), ('server', '450.61%'), ('awes
ome', '430.46%'), ('worst', '405.16%'), ('time', '398.41%'), ('people', '378.12%')]
Topic 8:
[('bad', '3170.64%'), ('experience', '1520.27%'), ('ok', '1021.76%'), ('seat', '354.45%'), ('loading', '336.74%'), ('pathetic', '316.42%'),
('available', '297.64%'), ('site', '290.05%'), ('performance', '285.95%'), ('system', '230.4%')]
Topic 9:
[('excellent', '1691.99%'), ('not', '1660.06%'), ('working', '1631.66%'), ('work', '931.86%'), ('properly', '815.66%'), ('useful', '807.7
3%'), ('open', '674.51%'), ('most', '367.57%'), ('time', '304.63%'), ('thank', '242.57%')]
```

Figure 11: Intertopic distance plot for TF-IDF Vectorizer.

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
21	TOPIC_0	Always hangs and lags often never work smoothly	always hang lag often never work smoothly	time, not, update, problem, opening, unable, always, worst, please, every
71	TOPIC_0	App crash too much Sometimes when I book tickets then app doesn't work it's need improvement	crash much sometimes work need improvement	time, not, update, problem, opening, unable, always, worst, please, every
110	TOPIC_0	Useless	useless	time, not, update, problem, opening, unable, always, worst, please, every
111	TOPIC_0	Indian government 🐷	government	time, not, update, problem, opening, unable, always, worst, please, every
134	TOPIC_0	Love it	love	time, not, update, problem, opening, unable, always, worst, please, every

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
11	TOPIC_1	Pls book to IRCTC app no extra charges ❤️🙏🚂🚃	no extra charge	worst, service, super, seen, one, india, used, difficult, happy, life
15	TOPIC_1	ALWAYS AMOUNT DEDUCTED FROM MY ACCOUNT BUT TICKETS NOT ISSUED....! MANY TIMES THIS HAPPENING..... CORRECT THIS PROBLEM....!	always amount deducted account not issued many time happening correct problem	worst, service, super, seen, one, india, used, difficult, happy, life
31	TOPIC_1	Failed no ticket booking.Also refund not received	failed no refund not received	worst, service, super, seen, one, india, used, difficult, happy, life
46	TOPIC_1	Super	super	worst, service, super, seen, one, india, used, difficult, happy, life
50	TOPIC_1	refund amount very low	refund amount low	worst, service, super, seen, one, india, used, difficult, happy, life

Figure 12: Sample documents reviewed for relevance (Count Vectorizer).

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
4	TOPIC_0	One of the pathetic app to use. Trying to re-download my ticket from app and every time there is an error.	one pathetic trying download every time error	time, not, update, problem, opening, unable, always, worst, please, every, issue, day, open, payment, error, showing, many, transaction, show, even
21	TOPIC_0	Always hangs and lags often never work smoothly	always hang lag often never work smoothly	time, not, update, problem, opening, unable, always, worst, please, every, issue, day, open, payment, error, showing, many, transaction, show, even
32	TOPIC_0	it is my MAJBOORI to ise this app otherwise i would have never installed it. It is torture to use this app.	never installed torture	time, not, update, problem, opening, unable, always, worst, please, every, issue, day, open, payment, error, showing, many, transaction, show, even
71	TOPIC_0	App crash too much Sometimes when I book tickets then app doesn't work it's need improvement	crash much sometimes work need improvement	time, not, update, problem, opening, unable, always, worst, please, every, issue, day, open, payment, error, showing, many, transaction, show, even
104	TOPIC_0	Getting an error "Unable to retrieve data, please try again"	getting error unable retrieve data please try	time, not, update, problem, opening, unable, always, worst, please, every, issue, day, open, payment, error, showing, many, transaction, show, even

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
46	TOPIC_1	Super	super	worst, service, super, seen, one, india, used, difficult, happy, life, bed, running, fantastic, shame, yet, platform, slowly, history, low, respond
66	TOPIC_1	One of the ghatia aap	one	worst, service, super, seen, one, india, used, difficult, happy, life, bed, running, fantastic, shame, yet, platform, slowly, history, low, respond
72	TOPIC_1	Worst app	worst	worst, service, super, seen, one, india, used, difficult, happy, life, bed, running, fantastic, shame, yet, platform, slowly, history, low, respond
148	TOPIC_1	Worst 👎services... Now	worst service	worst, service, super, seen, one, india, used, difficult, happy, life, bed, running, fantastic, shame, yet, platform, slowly, history, low, respond
152	TOPIC_1	Worst	worst	worst, service, super, seen, one, india, used, difficult, happy, life, bed, running, fantastic, shame, yet, platform, slowly, history, low, respond

Figure 13: Sample documents reviewed for relevance (TF-IDF Vectorizer).

3.5.5 Conclusion and Next Steps

With these results, there was scope for improvement. One key issue was that since data is noisy (which real-world data usually is), to handle it, we decided to change the probabilistic approach of getting topic and word distributions; we go by word embeddings to develop word context and similarities and perform topic mining on that. Hence, the following approach.

3.6 Approach 2: Word2Vec Embeddings

To try a word embedding approach in our data, we decided to make Word2Vec embeddings from our data after cleaning the data using the custom cleaning function developed for all models.

3.6.1 Text Preprocessing

We started to make a 100-dimensional word embedding for each word in our corpus/dictionary, and we plotted a t-SNE plot to visualize these vectors in a high-dimensional vector space into a 2-D dimensional space.

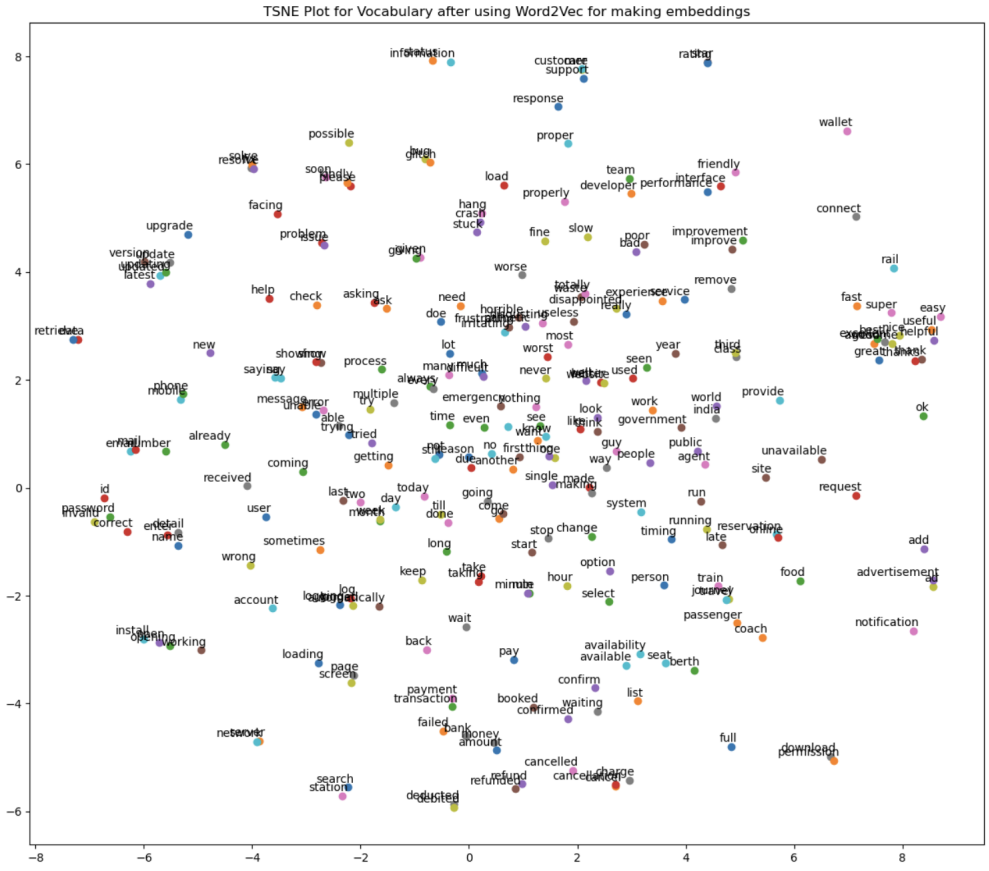

Figure 14: Word Embedding Plot for key words, showing semantic relationships in two-dimensional space.

We plotted only keywords in the image above for better visualization. However, in the project, we used all the words, not just those shown in the plot. With this, we embedded each word. The next step was to find

embedding for each document in the data. For this, we averaged words in the cleaned document text.

Also, we can see similar words being grouped nearby, for example: ”deducted” and ”debited.” However, we still see scope for improvement, which we will achieve by using higher-dimensional embeddings trained on larger datasets, specifically BERT embeddings in later approaches.

3.6.2 Modeling

Since we have real-valued embeddings, we used clustering-based modeling for this data. Moreover, the following is the result.

Model Evaluation:

- This involved silhouette scores, t-SNE plots (similar to intertopic distance in LDA), and human-based checking of crucial topic words and documents.
- **Silhouette Score:** 0.3124418 for 10 clusters, indicating moderate separation.

3.6.3 Evaluation Visuals

Figure 15: t-SNE Plot of Clustered Reviews using Word2Vec embeddings, showing semantic clustering of similar reviews.

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
0	TOPIC_3	Best	best	good, time, experience, money, slow, bad, problem, no, always, every, , opening, show, option, account, please, work, day, poor, error, need, train, try, showing, one, service, deducted, update, user, many
1	TOPIC_3	Good	good	good, time, experience, money, slow, bad, problem, no, always, every, , opening, show, option, account, please, work, day, poor, error, need, train, try, showing, one, service, deducted, update, user, many
2	TOPIC_3	Good 👍	good	good, time, experience, money, slow, bad, problem, no, always, every, , opening, show, option, account, please, work, day, poor, error, need, train, try, showing, one, service, deducted, update, user, many
3	TOPIC_3	There are good apps for booking tickets but a good internet is required to book instant tickets	good good internet required instant	good, time, experience, money, slow, bad, problem, no, always, every, , opening, show, option, account, please, work, day, poor, error, need, train, try, showing, one, service, deducted, update, user, many
4	TOPIC_3	One of the pathetic app to use. Trying to re-download my ticket from app and every time there is an error.	one pathetic trying download every time error	good, time, experience, money, slow, bad, problem, no, always, every, , opening, show, option, account, please, work, day, poor, error, need, train, try, showing, one, service, deducted, update, user, many

MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2

	MAIN_TOPIC	translated_content	translated_content_cleaned	review_category2
26812	TOPIC_5	After allowing all the require permission I still can't download my ticket	allowing require permission still download	not, worst, booked, time, nice, good, no, even, slow, always, experience, problem, unable, need, opening, poor, error, showing, try, , user, service, deducted, work, show, option, update, many, train, account
27247	TOPIC_5	While the ticket booking is seamless when using BHIM UPI, same can't be said about credit cards. Similarly, inability to print or save ticket to gallery because of non-existent storage permission is a bother.	seamless said credit card similarly inability print save gallery non existent storage permission bother	not, worst, booked, time, nice, good, no, even, slow, always, experience, problem, unable, need, opening, poor, error, showing, try, , user, service, deducted, work, show, option, update, many, train, account
29882	TOPIC_5	Storage permission	storage permission	not, worst, booked, time, nice, good, no, even, slow, always, experience, problem, unable, need, opening, poor, error, showing, try, , user, service, deducted, work, show, option, update, many, train, account

Figure 16: Manual validation of clustered words.

3.6.4 Conclusion and Next Steps

With these results, we identified a scope for improvement. We wanted to refine the quality of our embeddings since the image representations signal that heavily dissimilar words clustered. Therefore, before using BERT, we considered using a smaller-sized pre-trained model, like the Google News dataset, which is freely available. This approach minimizes the model size and aligns it with reduced computing requirements.

3.7 Approach 3: Google News Data Embeddings and LDA within Clustering

We used pre-trained Google News embeddings with higher levels of contextual information than the Word2Vec embeddings specifically tailored to our dataset. We first performed clustering on the data, and then apply LDA within each of these clusters for better topic separation of dissimilar topics (different clusters) and finer topic differentiation (within each cluster using LDA.

3.7.1 Text Preprocessing

We applied our cleaning function to the text and preprocessed it before generating embeddings, similar to what we would do with the data before feeding it into any model for training. We are using Google News embeddings, similar to the Word2Vec example above, but with pre-trained vector embeddings.

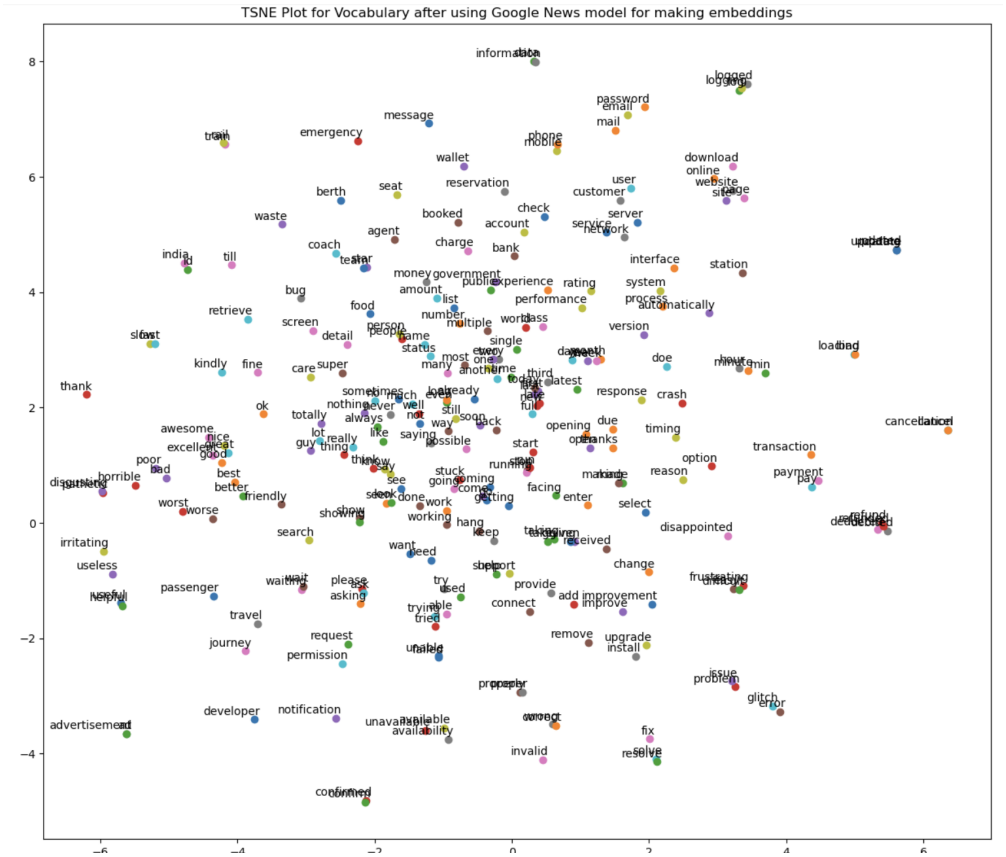

Figure 17: t-SNE Plot of Word Embeddings using Google News data, showing clusters of semantically similar words.

This visualization allowed us to confirm that the Google News embeddings effectively grouped semantically similar words closer together.

3.7.2 Modeling

In this approach, we first clustered data by a fixed number of clusters using the embeddings from Google News. For deeper patterns within these clusters, we applied LDA within each cluster and measured overall effectiveness of topics obtained.

3.7.3 Model Evaluation Metrics:

- **Silhouette Score:** 0.3086092 for clustering, indicating moderate separability.
- **Per-Cluster Evaluation:** We calculated perplexity and log likelihood scores for each cluster to evaluate the coherence of topics within each.

3.7.4 Cluster Evaluation Results

- **Cluster 0:** Log Likelihood: $-4,770.70$, Perplexity: 10.82

- **Cluster 1:** Log Likelihood: $-357,353.30$, Perplexity: 229.64
- **Cluster 2:** Log Likelihood: $-3,096,821.72$, Perplexity: 708.46
- **Cluster 3:** Log Likelihood: $-15,500.84$, Perplexity: 23.13
- **Cluster 4:** Log Likelihood: $-22,880.02$, Perplexity: 5.60
- **Cluster 5:** Log Likelihood: $-15,997.03$, Perplexity: 17.75
- **Cluster 6:** Log Likelihood: $-2,051.39$, Perplexity: 7.36
- **Cluster 7:** Log Likelihood: $-6,677.94$, Perplexity: 11.95
- **Cluster 8:** Log Likelihood: $-2,387.10$, Perplexity: 7.50
- **Cluster 9:** Log Likelihood: $-11,312.02$, Perplexity: 7.31

3.7.5 Words within each topic of each Cluster

To validate the coherence of the topics within each cluster, we manually reviewed words associated with each topic and provided a summary table below.

Cluster	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
0	best, one, nice, option, wish	best, good, easy, india, performance	best, experience, reservation, travel, helpful	best, rail, online, site, seller	best, train, travelling, smooth, hotel
1	not, able, bad, like, helpful	not, working, slow, properly, opening	not, work, open, time, never	good, experience, service, nice, better	useful, easy, need, useless, lot
2	not, time, update, open, opening	time, payment, not, unable, worst	user, not, ad, password, id	worst, service, no, one, poor	not, money, booked, train, no
3	good, stupid, irritating, awful, download	dumb, horrendous, unfriendly, disastrous, bad	poor, bad, service, terrible, work	bad, experience, slow, performance, useless	bad, pathetic, not, disgusting, worst
4	good, enough, decent, enquiry, quite	good, better, great, best, fine	good, easy, helpful, useful, happy	good, job, thank, well, bad	good, luck, like, really, overall
5	worst, seen, india, train, work	worst, used, world, one, slow	worst, not, performance, worse, cancelled	worst, time, server, most, download	worst, experience, service, never, site
6	shit, ad, improved, ok, cant	ok, good, cant, thanks, fine	gee, ki, ok, cant, sir	okay, yeah, oh, ji, well	ok, yes, overall, oh, hm
7	superb, amazing, extraordinary, easy, splendid	wonderful, thanks, helpful, impressive, unbeatable	fantastic, experience, brilliant, useful, reservation	great, service, outstanding, work, super	excellent, good, performance, perfect, fabulous
8	super, good, experience, excited, fast	id, super, cute, sir, fastest	super, product, cute, sir, fastest	super, slow, time, used, performance	super, simply, india, ok, se
9	nice, useful, pic, happy, easily	awesome, thanks, ok, fast, summer	nice, helpful, good, interesting, liked	beautiful, lovely, amazing, different, beer	nice, good, easy, super, wow

Table 1: Summary of top words by cluster and topic

3.7.6 Evaluation Visuals

	clusters	topics	words_in_topic	content	translated_content	output
0	9	1	best, experience, good, one, easy	Best	Best	cluster: 0topic: 2 with probability=0.5932227644461392
1	1	1	good, slow, nice, enough, okay	Good	Good	cluster: 3topic: 0 with probability=0.5999980298689461
2	1	1	good, slow, nice, enough, okay	Good 👍	Good 👍	cluster: 3topic: 0 with probability=0.5999980298689461
3	6	1	time, worst, not, server, every	Ticket book karne ke liye achha apps hai lekin tatkal tickets book karne ke liye achha net ki jarurat hai	There are good apps for booking tickets but a good internet is required to book instant tickets	cluster: 2topic: 0 with probability=0.5851719955986292
4	6	1	time, worst, not, server, every	One of the pathetic app to use. Trying to re-download my ticket from app and every time there is an error.	One of the pathetic app to use. Trying to re-download my ticket from app and every time there is an error.	cluster: 2topic: 1 with probability=0.8962765201387567
...	...	...	...	...	...	...
64240	9	0	best, think, absolutely, amazing, public	make easiest usefull.	make easiest usefull.	cluster: 1topic: 4 with probability=0.5999920324792262
65316	8	3	ji, gee, ok, sir, well	Gigh	Gee	cluster: 2topic: 0 with probability=0.2
75399	8	3	ji, gee, ok, sir, well	Ok hai bhiji	ok hai bhi ji	cluster: 2topic: 3 with probability=0.7294763986572755
78450	2	1	super, star, travelling, personal, sir	super travelling app that's	super travelling app that's	cluster: 2topic: 3 with probability=0.5085413954712523
90414	2	2	se, super, like, speed, sir	Super app for ticket booking ❤Se.👌👌👌	Super app for ticket booking ❤Se.👌👌👌	cluster: 2topic: 3 with probability=0.733201700969445

Figure 18: Sample of human-validated reviews for relevance within clusters.

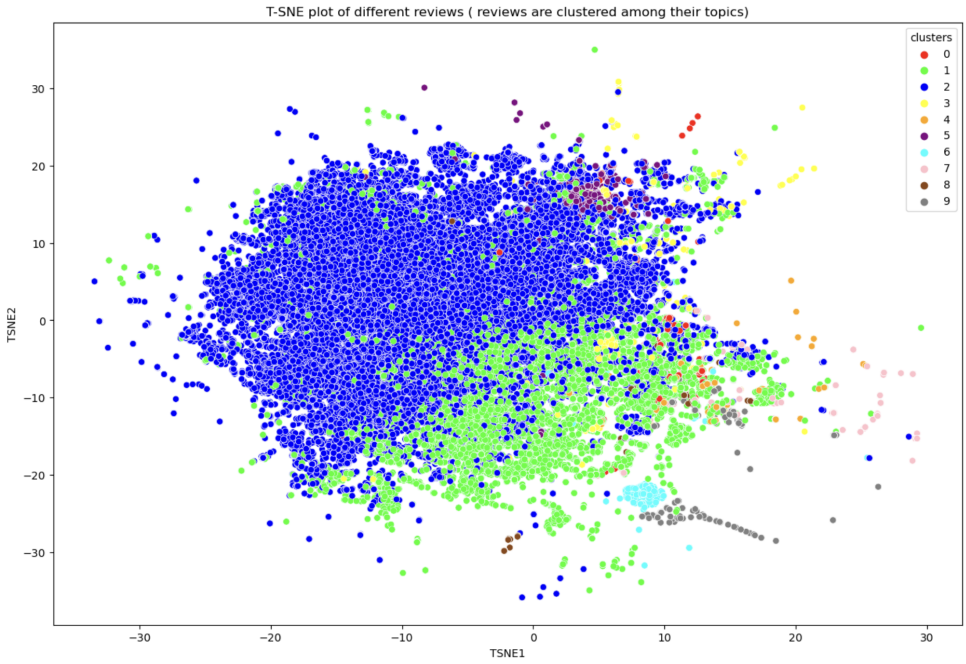

Figure 19: t-SNE Plot of clustered reviews using Google News embeddings.

3.7.7 Conclusion and Next Steps

This confirmed our hypothesis that the clustering technique coupled with topic modeling within clusters would reveal meaningful topic structures otherwise would be complex due to overlap in topics and very irregular shapes of clusters in t-SNE plots. It showed there was still room for improvement both in embedding quality and in the clustering method.

We were motivated by the usage of BERT embeddings with hierarchical and density-based clustering because it will cover up the shortcomings of the above system, considering the fact that BERT's pre-trained language model captures better contextual information, which we hypothesize improves both topic separation/grouping and cluster quality.

3.8 Approach 4: BERT Embeddings and Clustering-Based Topic Mining

We have revised our final approach by using BERT embeddings for the tasks of representing the subtle semantic relationships existing in user reviews. We used a framework called BERTopic for building a topic mining model using hierarchical density-based clustering and UMAP model.

3.8.1 Text Preprocessing

We used the same cleaning function on text mentioned in common preprocessing steps.

3.8.2 Modeling

We applied UMAP for dimensionality reduction to reduce the high-dimensional BERT embeddings, thus it helped retain essential context in the data making clustering feasible. Then we applied the hierarchical density-based clustering algorithm HDBSCAN on the reduced data, allowing clusters of complex shapes and a robust noise handling.

Using this pipeline, we arrived at final model that was generating a number of unique topics throughout the review data. Using Class-based TF-IDF (C-TF-IDF) for each discovered cluster, we were able to attain the most significant words that correspond to each cluster found and more direct focus to the unique terms in each topic. After manual validation, we have condensed the topics down to 66 themes or topics.

3.8.3 Topic Visualization and Evaluation

For clustering we got Silhouette score of 0.21864341

Note: The plots below use the default BERTopic labels, based on the most frequent words in a given topic. We mapped these to human-readable labels in order for the data to be easier to interpret.

- **Intertopic Distance Plot:** Visualizes topic distances based on C-TF-IDF scores, highlighting the distribution/overlap and similarity between topics.
  - Intertopic Distance Plot
- **Topic Similarity Plot:** Shows similarity across the top 20 topics for a concise view. Clearly we are able to get distinct topics with hint of similarity on close clustered topics.

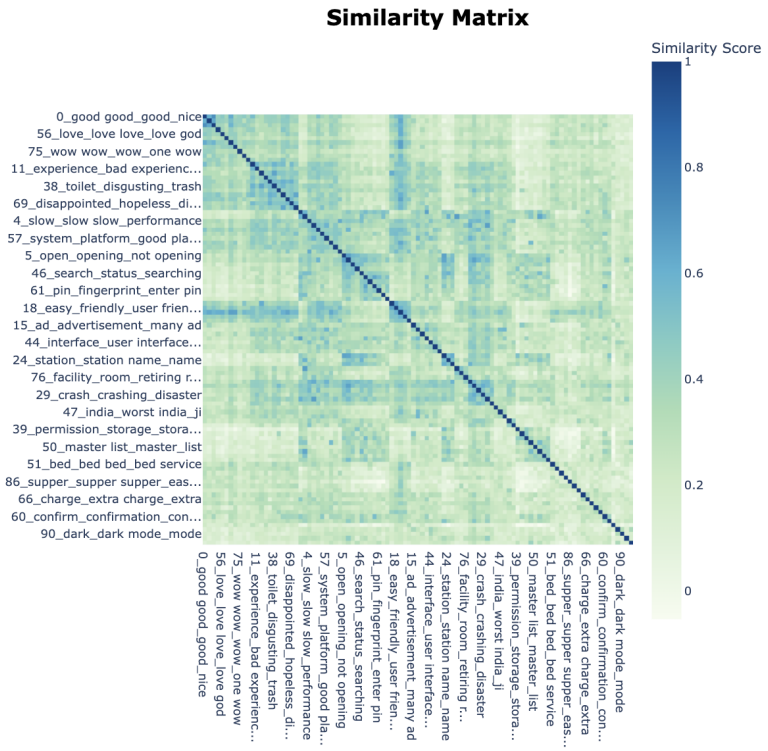

Figure 20: Topic Similarity Plot showing the top 20 topics.

- **Word Frequency Plots:** Sample word frequency plots for key topics, providing insights into commonly discussed terms.

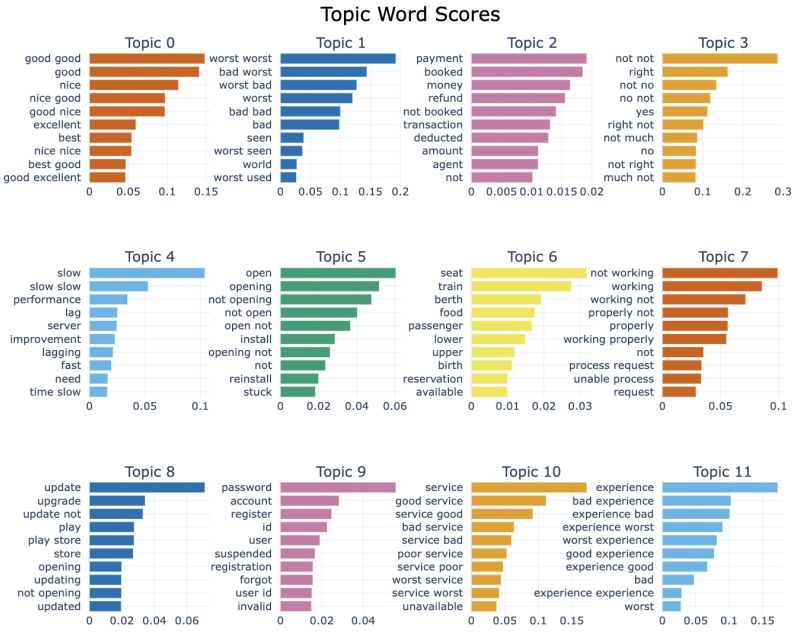

Figure 21: Word Frequency Plot for few sample topics.

Note: All topic plots are saved in the accompanying Jupyter Notebook for full reference, to keep the

report concise.

- **Topic Hierarchy Plot:** A hierarchical view of the top 40 topics, summarizing related themes.

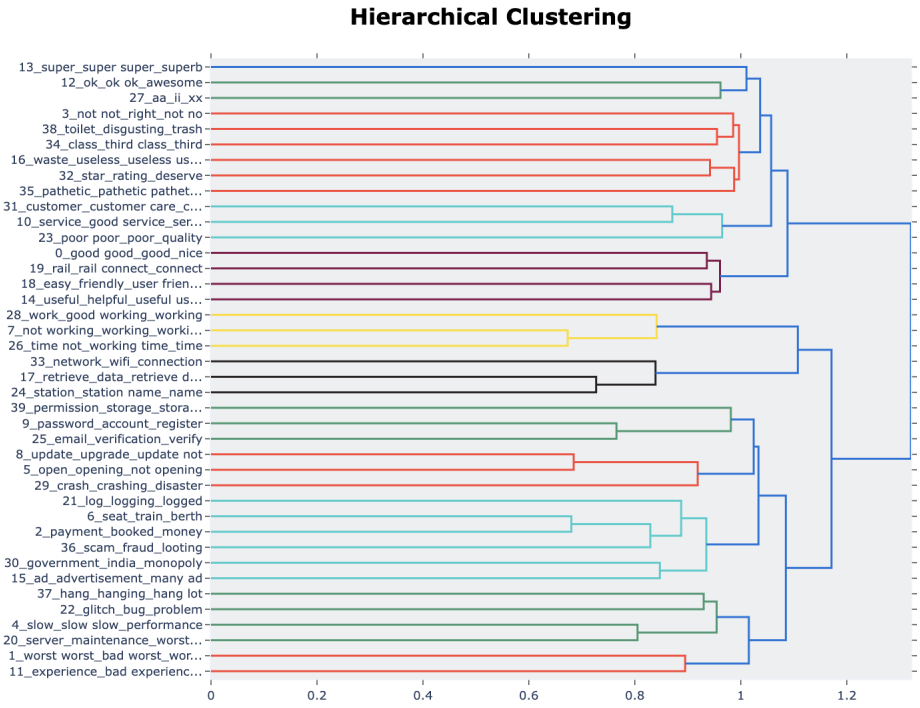

Figure 22: Topic Hierarchy Plot showing the top 40 topics.

- **Human-Readable Topic Table:** A table with manually labeled topic names for ease of interpretation.

Topic ID	Representative Words	Human-Readable Label
0	train, booking, app, time	Booking Process
1	service, issue, response, support	Customer Support
2	payment, failed, refund, money	Payment Issues
...	...	... please refer to the jupyter notebook for complete list of unique 66 topics

Table 2: Human-Readable Topic Table summarizing key themes in the data.

- **t-SNE Plot of Topic Embeddings:** Shows clusters in two-dimensional space for visualization of topic spread and density.

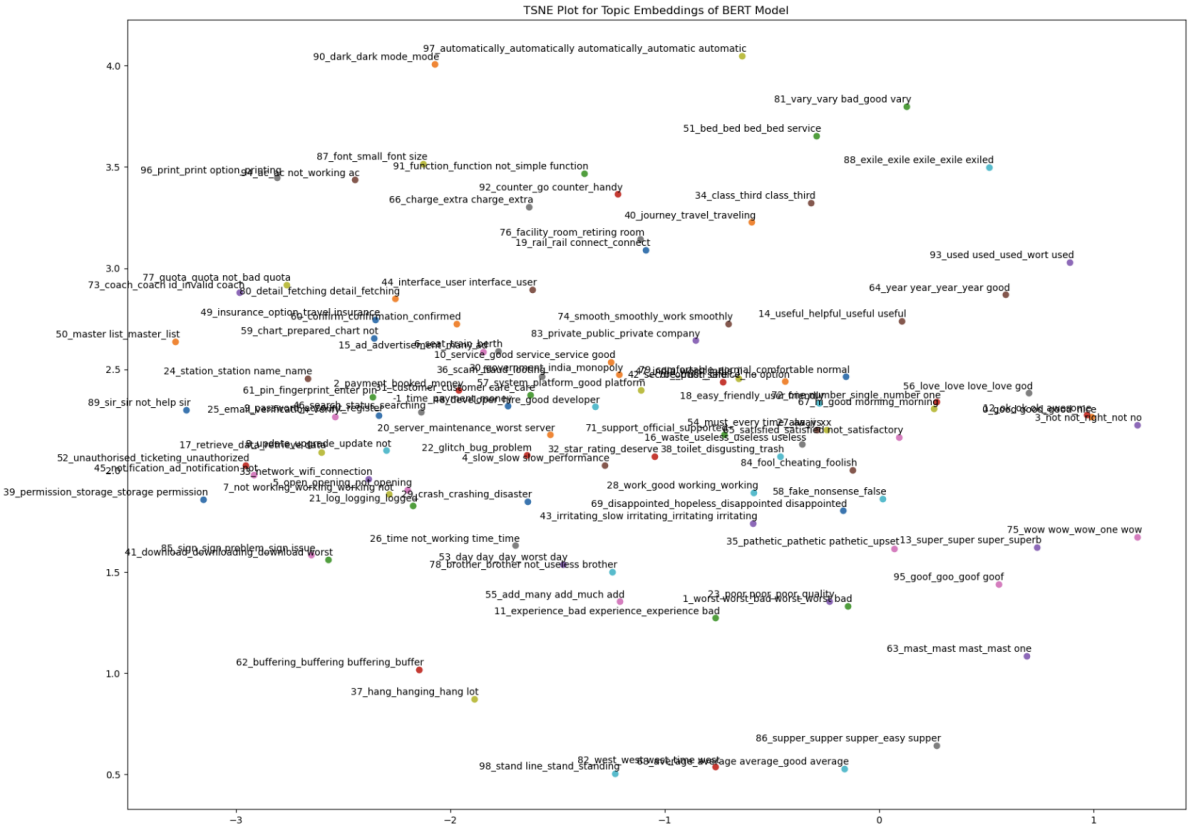

Figure 23: t-SNE Plot of Topic Embeddings for BERTopic model.

3.8.4 Conclusion and Model Finalization

This model categories well-defined topics that enlists the core issues and strengths raised by the users in their reviews. This showed that the model indeed met our objectives, like unfolding to unique topics, thus giving a comprehensive understanding of user sentiment and context plus pointing out actionable areas to improve on for the IRCTC team.

We have uploaded the model (details mentioned in the README.md file) in the interest of size, so you can easily access it. The final model is robust for ongoing analysis; this would enable IRCTC to monitor the user feedback and continuously enhance its app experience.

4 Sentiment Analysis

For additional insights into the data, sentiment analysis was performed on the review data. Text model alone might not be sufficient to capture the best inference of the topic being discussed. Hence, sentiment analysis will help us classify the reviews more accurately.

4.1 Sentiment Analysis Using Gensim

Initially, "Gensim" framework was used to perform sentiment analysis. Various type of analysis was done on the data including but not limited to mapping numbered rating to the reviews, plotting word cloud of positive and negative reviews, etc. We calculated polarity values of reviews using Gensim, and used it classify into negative vs positive reviews

4.1.1 Word Clouds

Word clouds were generated for positive and negative reviews to observe the dominant words associated with both categories.

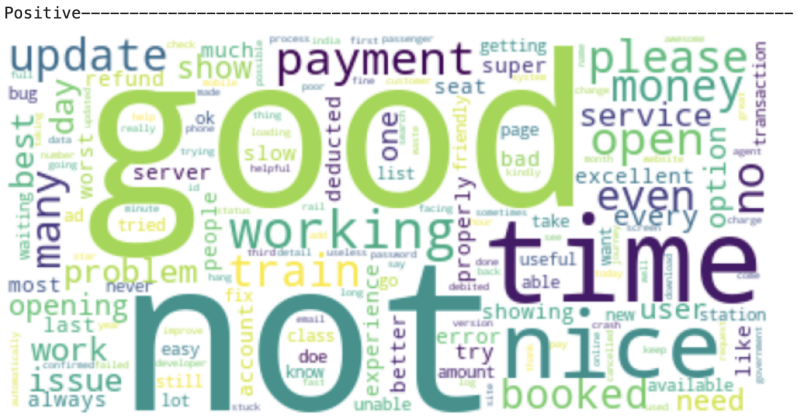

Figure 24: Word Cloud for Positive Reviews (Gensim).

Figure 25: Word Cloud for Negative Reviews (Gensim).

4.1.2 Polarity vs. Rating Table

Using these insights, ratings mapped from review were compared to actual star ratings given by users. It was found that star rating was very noisy as many users gave bad reviews but good star rating or vice versa. This kind of noise is common with real data. Which is why our model is particularly useful for giving more refined analysis of feedback data to our clients. Clients can leverage this insights to make more tactical business decisions.

score	1	2	3	4	5
sentiment_outcome					
Negative	27002.0	1852.0	863.0	380.0	471.0
Positive	26077.0	3932.0	4095.0	6552.0	28776.0

Figure 26: Rating vs Polarity

4.2 BERT-Based Sentiment Analysis

To achieve better sentiment analysis BERT (Bidirectional Encoder Representations from Transformers) embedding was used. Since BERT captures syntactic as well as semantic information of large corpus of data, it provides us with more accurate sentiment analysis when compared to custom embeddings or other methods like isolated word frequencies.

	content	sentiment	sentiment_score	Our_Label_of_Topic	bert_topic
0	Best	POSITIVE	0.999794	Positive Reviews	0
3	Ticket book karne ke liye achha apps hai lekin tatkal tickets book karne ke liye achha net ki jarurat hai	POSITIVE	0.999566	General Review about online platform	20
4	One of the pathetic app to use. Trying to re-download my ticket from app and every time there is an error.	NEGATIVE	0.999728	Payment, Money and Slow-loading Related Issues	-1
5	Walaikum assalam	POSITIVE	0.748121	Vague	3
6	Very comfortable	POSITIVE	0.999855	General experience/comfort related reviews	79
...	...	...	...	...	...
7425	तत्काल बुकिंग का समय चुटिया बनाता है लुगो को 🤦🤬	NEGATIVE	0.999520	Vague	84
7746	Best to use Railway counter... Worst app	NEGATIVE	0.995232	Reviews comparing counter ticket vs app	92
7981	Why you have ads? Allow to change/increase the font size.	POSITIVE	0.835545	Font Size related reviews	87
8735	Unable to print tkt	NEGATIVE	0.999579	Print feature related reviews	96
11286	Out standing	POSITIVE	0.999708	General Reviews	98

Figure 27: Sentiment Analysis using BERT

4.3 Sentiment Distribution Analysis

Sentiment distribution across topics mappings were plotted to visualize the data and observe the trends.

4.3.1 Average Sentiment by Topic Plot

Following plots display the average sentiment for each major topic, giving a clear view of which topics regarding the app are viewed positively and which ones are viewed negatively.

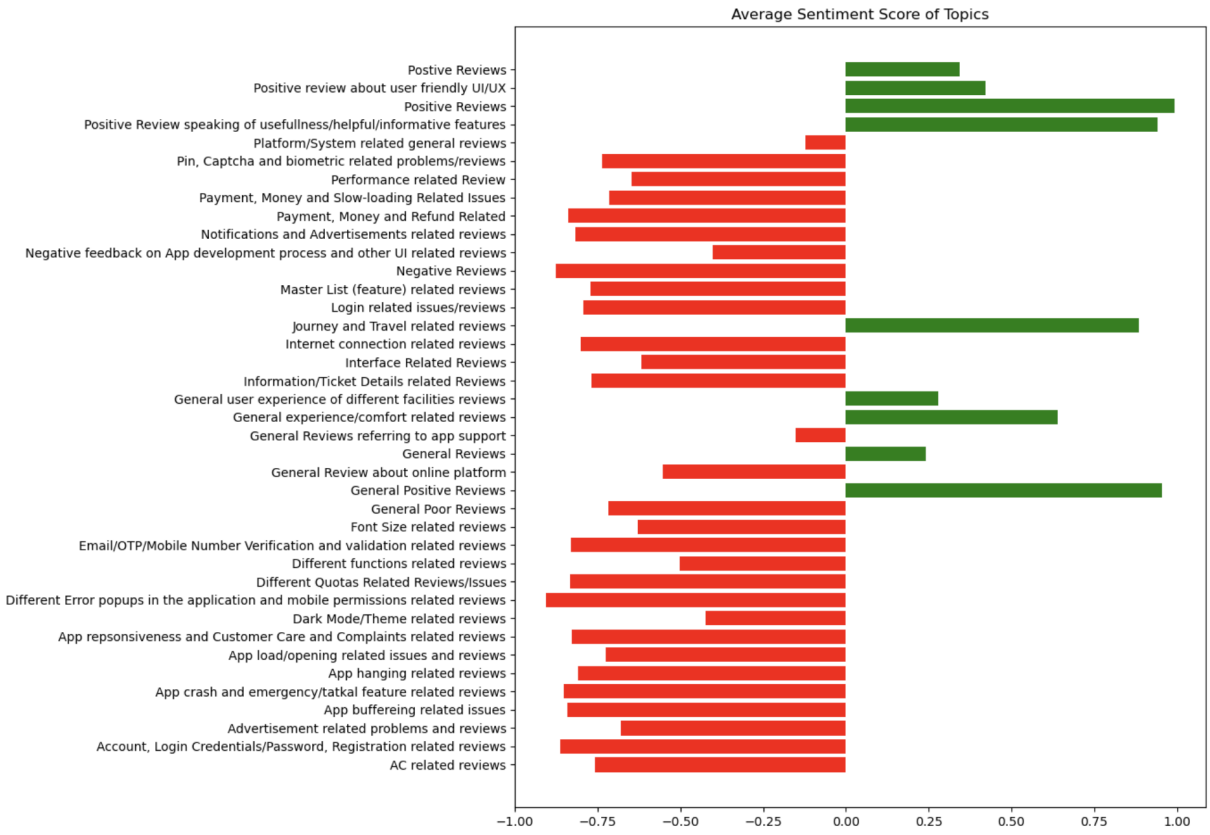

Figure 28: Average Sentiment Scores for Different Topics from the Final BERT Model- set 1

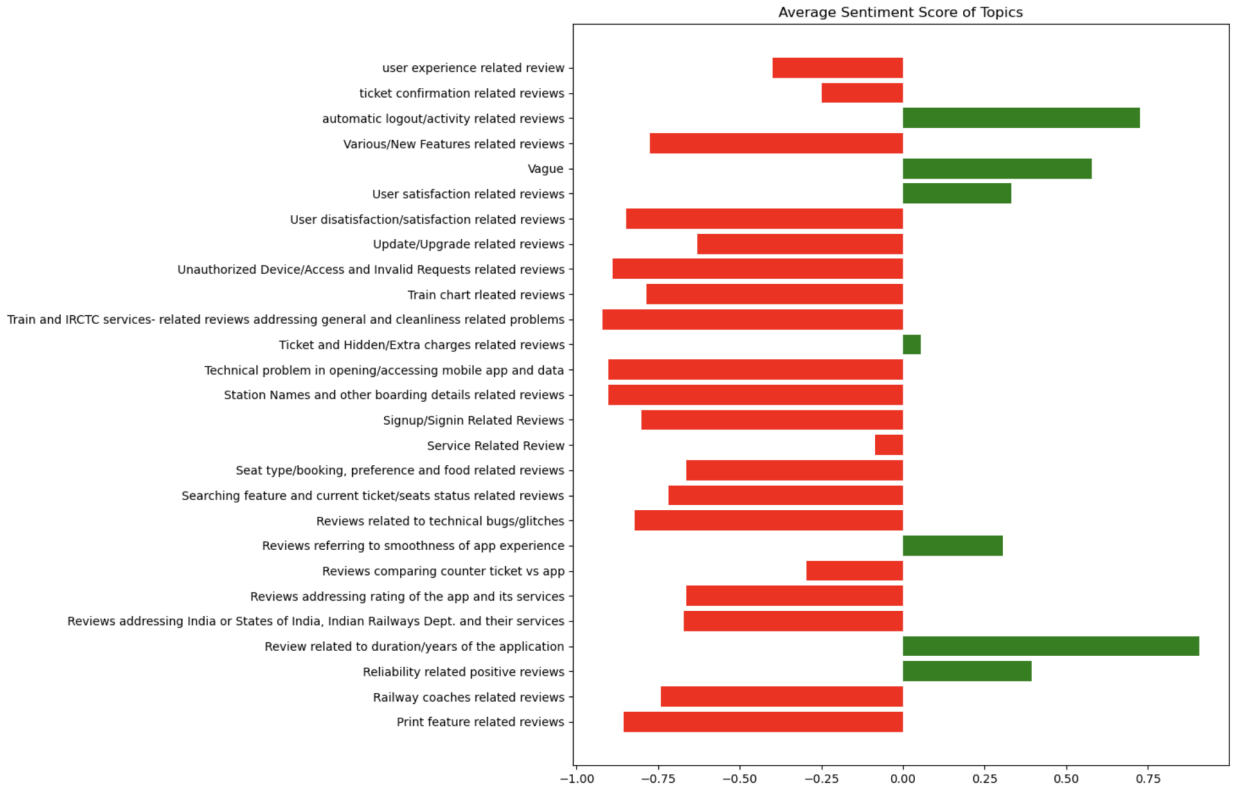

Figure 29: Average Sentiment Scores for Different Topics from the Final BERT Model- set 2

4.3.2 Sentiment Over Time Plot

Sentiment trends over time were examined to track shift in user sentiment. This gave us important insights into topics like user satisfaction, feature reception like if any particular app update resulted in spike of any particular issue at that time leading to any dips in the plot, etc.

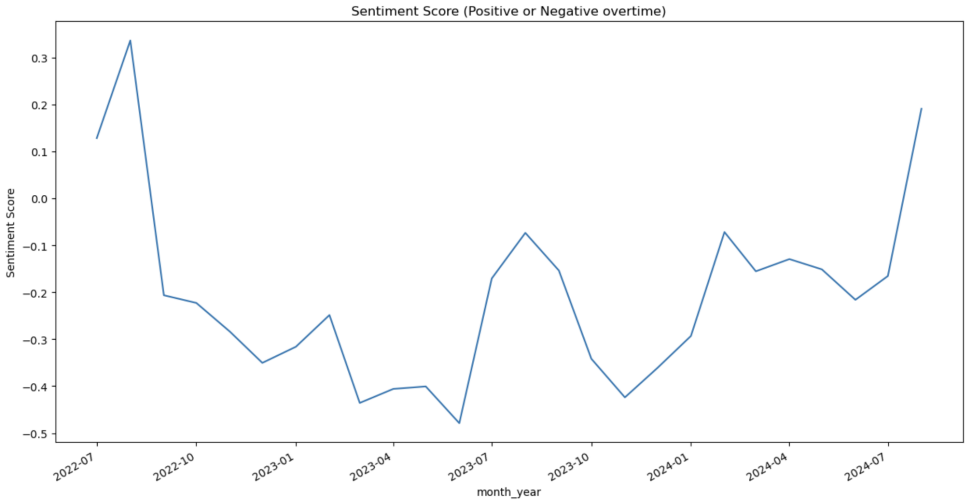

Figure 30: Sentiment over time, showing fluctuations in sentiment potentially linked to app updates or changes.

4.4 Conclusion

Sentiment analysis using Gensim and BERT gave a comprehensive view of user reviews that goes beyond text models. The Gensim word cloud gives a quick high-level visualization of common themes. Polarity vs Rating analysis showed inconsistencies between feedback reviews and star ratings. The BERT model added depth to analysis by capturing contextual information from reviews. This allowed us more accurate classification. Analysis of various topics over time gave us further insights into trends of user satisfaction levels. Together, these methods offers data driven and actionable insights to improve decision making for product improvements.

5 Results, Conclusions and Future Directions

In each of our Models, we have given details of Results, isights drawn from that approach and what future directions can be done for each of those approaches. Further, this section will give a summary of our analysis. It includes the hierarchy of topics, time series plots for key topics, and inferences drawn from our analysis. We have also added suggestions for IRCTC to enhance their app.

5.1 Hierarchy of Topics

Based on our advisors suggestions, we structured the topics into a hierarchy. This helps to interpret user concerns by grouping related themes and helps with prioritizing issues based on user reviews. Hierarchy of topics also helps cut down time taken and effort of going through each individual low level topics. Full Level 3 Grouping please refer to Jupyter notebook as its very long

Level 1 Grouping	Level 2 Grouping
Ambiguous or Vague	Ambiguous or Vague
General Reviews	Service Experience, User Experience, General Review, Platform Experience, Travel Experience, UI/UX Experience, Features and Usefulness, User Satisfaction, Booking and Travel Features
Issues	Payment and Performance Issues, Performance, App Loading and Access, Booking and Travel Features, Update and Maintenance, Account and Authentication, Advertisements, Technical Issues, Customer Service, Service Experience, UI/UX Issues
Negative Feedback	General Negative Feedback
Positive Feedback	General Positive Feedback, Features and Usefulness, UI/UX Experience, Reliability

5.2 Time Series Analysis of Key Topics

Trends of topics were analyzed over a time period to observe the changes in user sentiment and engagement. Following plots (7 in each) illustrate these trends.

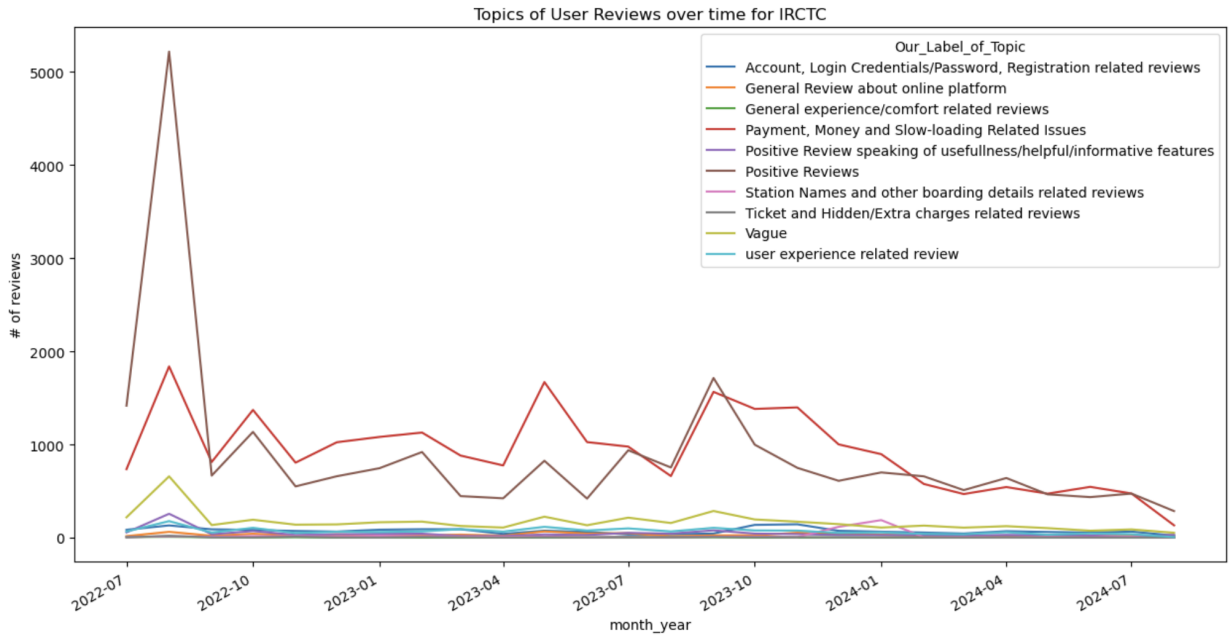

Figure 31: Time series plot for Set 1

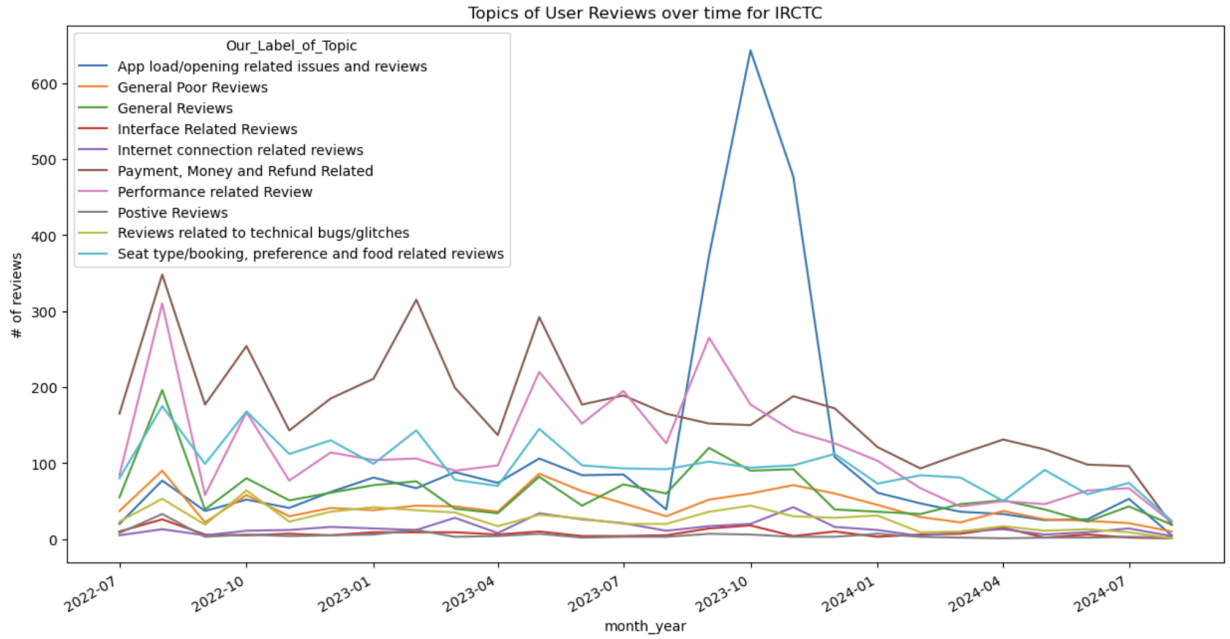

Figure 32: Time series plot for Set 2

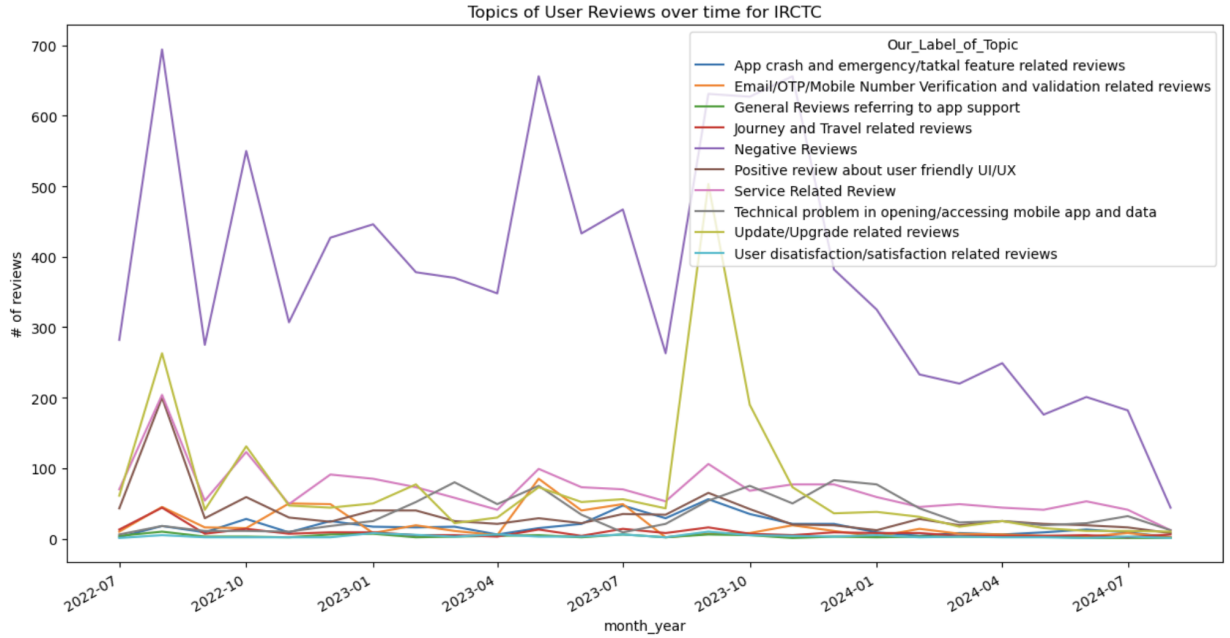

Figure 33: Time series plot for Set 3

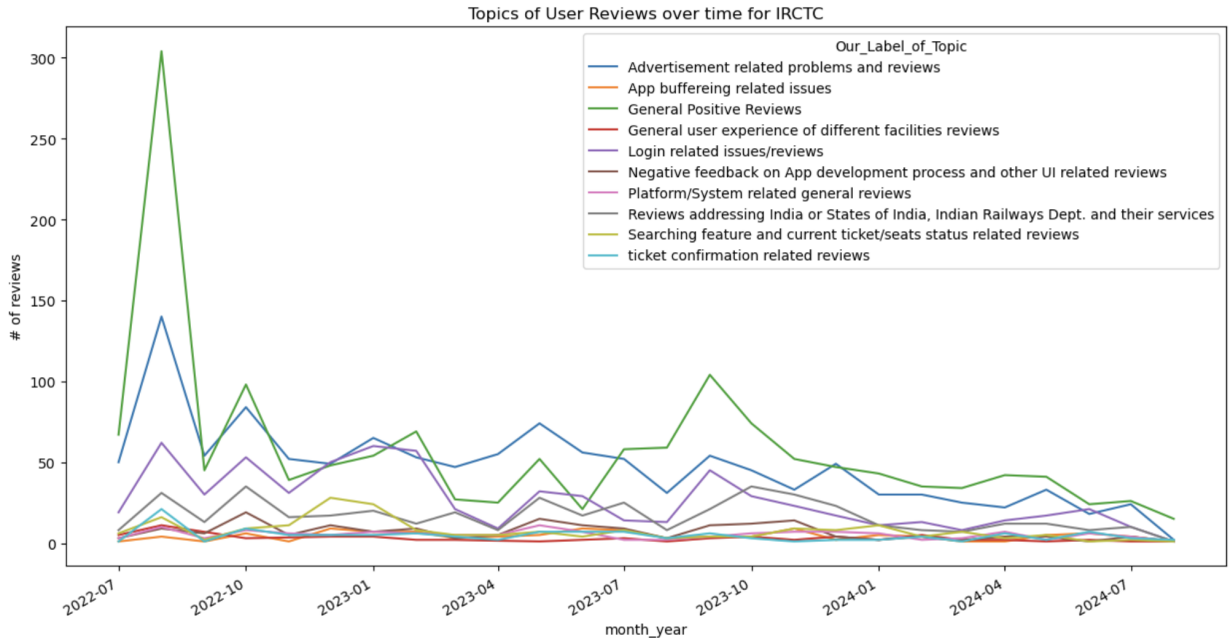

Figure 34: Time series plot for Set 4

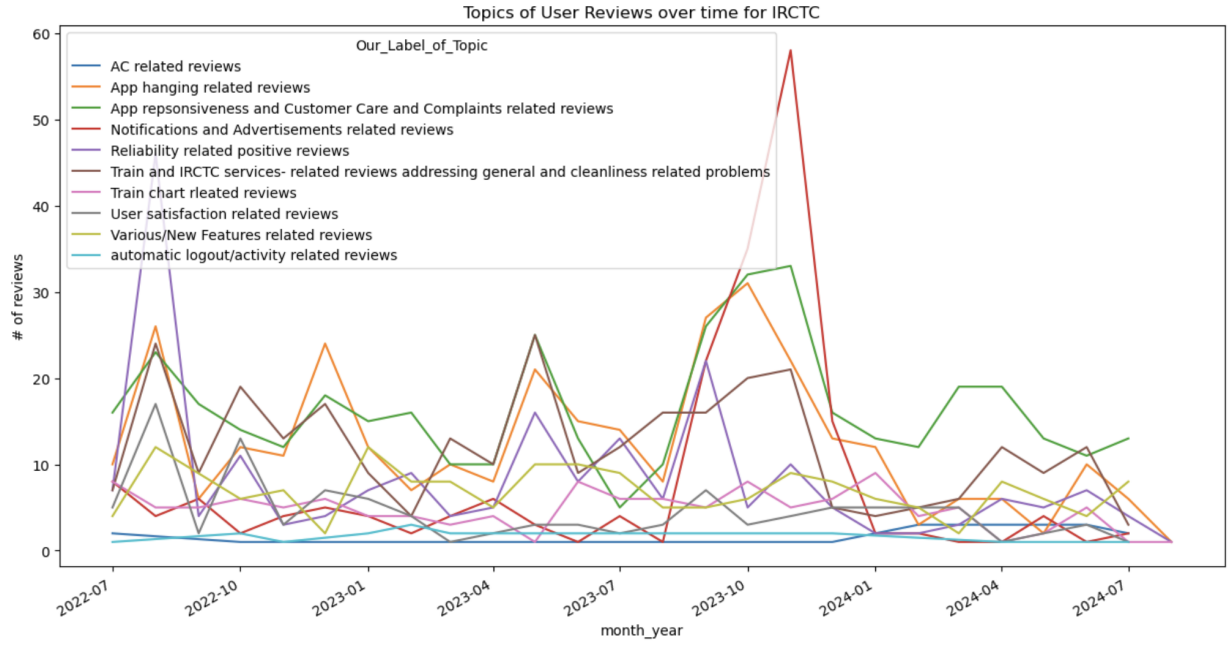

Figure 35: Time series plot for Set 5

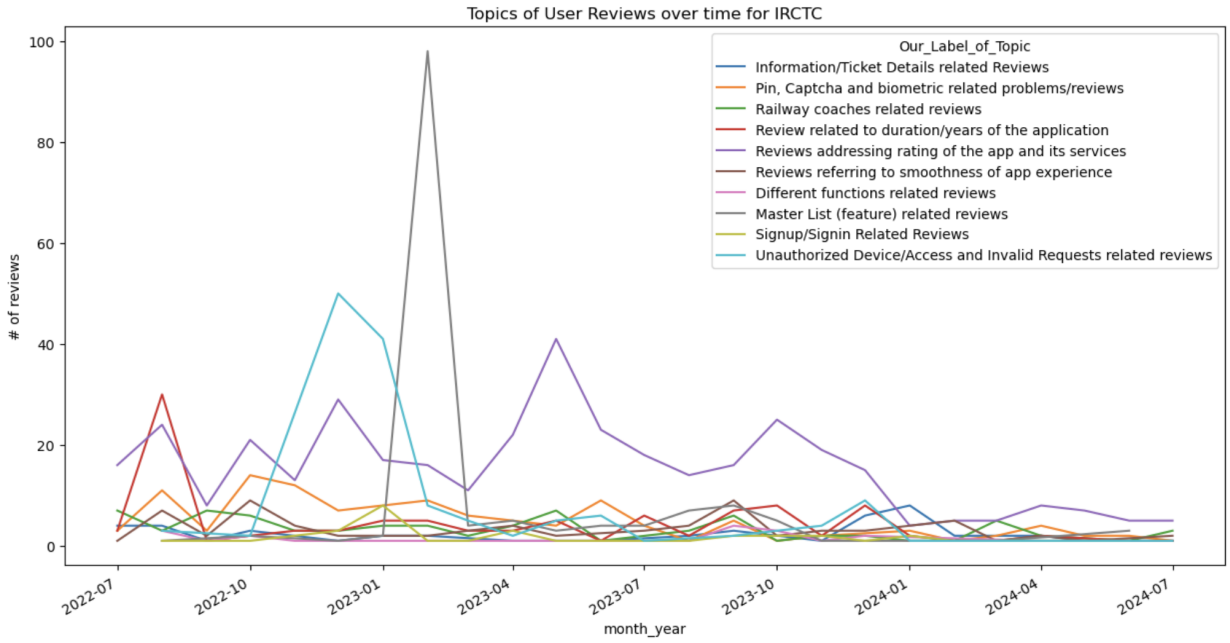

Figure 36: Time series plot for Set 6

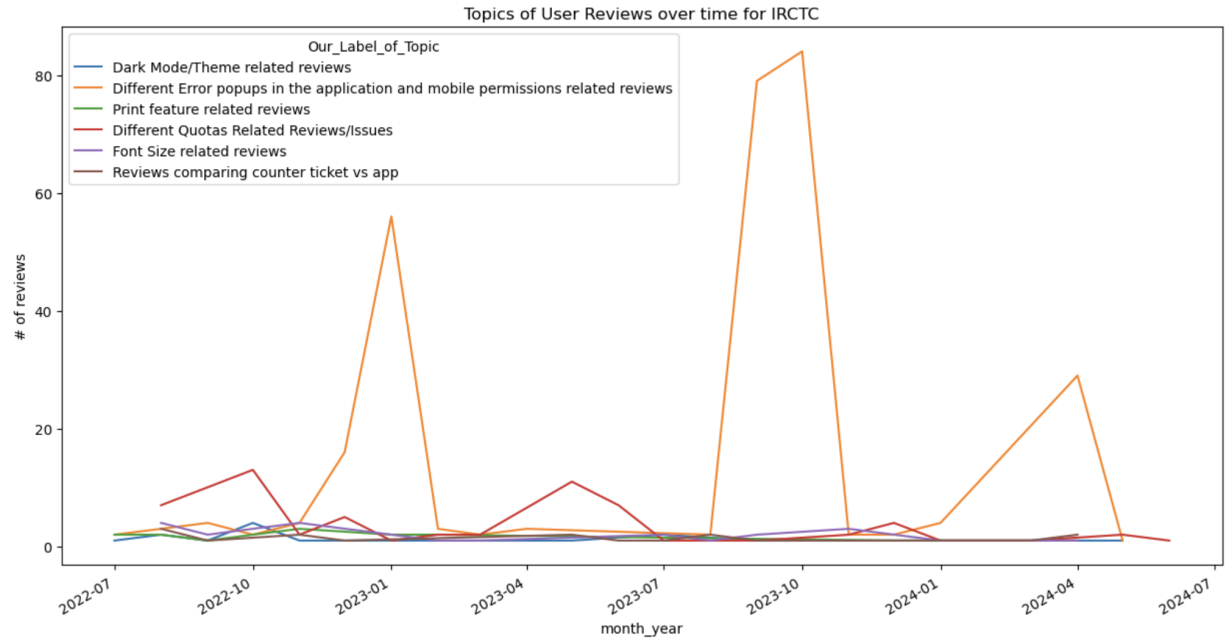

Figure 37: Time series plot for Set 7

5.3 Inferences from the Plots

Each time series plot gives particular insights into prevalence of issues faced by users over time.

- **Payment Concerns:** The frequency at which complaints regarding payments is received, especially such complaints being raised about delayed refund and transactions, suggests that the payment mechanism needs to be revamped. Improving payment infrastructure and communication with payment aggregators can be part of the solutions to these problems.

- **Booking Process:** There seem to be book-related problems review spikes coinciding with the holidays, when user traffic is high. This suggests improving the booking infrastructure at IRCTC may lead to elevation of user experience. This is backed by number of such reviews under the related topic seen in our topic mining results

- **App Performance:** App performance is consistently complained about in reviews. This points to flaws in App optimization. In addition to improving booking infrastructure, improving App optimization would be a key step to take, as most of the reviews spoke about app bugs, crashing and updates. Refer to the topics count list in the jupyter notebook.

- **Customer Service:** App disruptions mostly result in high customer service-related reviews, pointing to the dissatisfaction consumers have with response time along with the quality of the responses provided. This shows that improving customer support and response time are positive areas toward consumer satisfaction.

- **User Interface:** UI related reviews shows appreciation of simple and intuitive interface in users. Making minor improvements while maintaining current UI might be the key here.

- **General Feedback:** General sentiments captured under this topic shows a correlation between positive reviews and improvement in booking efficiency and app performance.

- **Pricing Concerns:** This topic captures complaints about pricing. These reviews often follow fare increase or addition in charges. Transparent communication by IRCTC regarding charge hikes might help reduction in number of such feedback.

5.4 Overall Remarks and Recommendations

Our final model, which uses BERT embeddings and hierarchical clustering, provides accurate and detailed insights into user experience on the IRCTC app. Few takeaways and recommendations are given below.

- **User-Centric Recommendations:** Sentiment analysis and topic mining give several user-centric improvements. These improvements include refining the booking process, improving payment reliability and optimizing App.

- **Targeted Improvements:** Improving App Optimization including App loading time, payment systems and customer response could be a good focus point for IRCTC.

5.5 Submission and Data Usage

- **Data & Scripts:** Project submission includes all Data and Scripts used. The trained model is available at the mentioned link in README.md file.

6 Team Contributions

Aditya and Vraj worked on making review scraper and data procurement. Aryan worked on methodology and building of topic mining and sentiment analysis models. Harsh worked on making demo website and text cleaning/verification tasks.

References

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *NLTK: The Natural Language Toolkit*. O'Reilly Media.

[Chollet et al., 2015] Chollet, F. et al. (2015). *Keras: The Python Deep Learning library*. GitHub repository, https://github.com/keras-team/keras.

[Grootendorst, 2020] Grootendorst, M. (2020). *BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling*. GitHub repository, https://github.com/MaartenGr/BERTopic.

[Loria, 2018] Loria, S. (2018). *TextBlob: Simplified Text Processing*. GitHub repository, https://github.com/sloria/TextBlob.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). *PyTorch: An Open Source Machine Learning Framework*. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[Sievert and Shirley, 2014] Sievert, C. and Shirley, K. (2014). *pyLDAvis: Python Library for Interactive Topic Model Visualization*. Journal of Open Source Software.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). *Transformers: State-of-the-Art Natural Language Processing for Pytorch and TensorFlow 2.0*. GitHub repository, https://github.com/huggingface/transformers.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). *Gensim: Topic Modelling for Humans*. Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks.

[Grootendorst, 2020] [Řehůřek and Sojka, 2010] [Chollet et al., 2015] [Bird et al., 2009] [Pedregosa et al., 2011] [Loria, 2018] [Paszke et al., 2019] [Wolf et al., 2020] [Sievert and Shirley, 2014]